

Bilingual Pivotingによる言い換え獲得の 相互情報量に基づく一般化

梶原 智之^{1,a)} 小町 守^{1,b)} 持橋 大地^{2,c)}

概要: 大規模な語彙的言い換え知識である PPDB が多くの自然言語処理応用タスクで利用されている。PPDB は Bilingual Pivoting と呼ばれる対訳コーパスを用いた単語アライメントによって言い換え対を獲得するが、大規模な反面、アライメント誤りに起因するノイズが多い。そこで本研究では、言い換えらしさを重み付き相互情報量によって定義し、Bilingual Pivoting によって得られる言い換え対を単言語コーパスから得られる分布類似度を用いてリランキングすることでノイズを軽減する。我々の提案手法は、相互情報量の低頻度問題に対処する重み付き相互情報量を言い換え獲得のために拡張するものである。分布類似度には Bilingual Pivoting のような関連のない語句によるノイズは起こりにくく、Bilingual Pivoting には分布類似度のような反義語や類義語によるノイズは起こりにくいため、これらを同時に考慮する提案手法は従来の言い換え獲得手法の欠点を補い合うことによる高い頑健性が期待できる。実験の結果、我々が提案する相互情報量に基づく言い換え獲得手法は、MRR と MAP の両方で Bilingual Pivoting を改善できた。

1. はじめに

情報検索や質問応答、機械翻訳の前編集やテキスト平易化など、多くの自然言語処理応用タスクにおいて、言い換えによる柔軟な言語理解が重要である。近年、大規模な語彙的言い換え知識である PPDB [1,2] が公開され、意味的文間類似度 [3] や単語分散表現の学習 [4] などに利用されている。PPDB は図 1 の Bilingual Pivoting [5] と呼ばれる対訳コーパスを用いた単語アライメントによって言い換え対を獲得するが、この単語アライメントには誤りも多い。

本研究では、Bilingual Pivoting によって対訳コーパスから得られる言い換え対を単言語コーパスから得られる情報を用いてリランキングし、言い換えらしさを相互情報量によって定義する。そして、相互情報量の低頻度問題に対処する LocalPMI [6] を言い換え獲得タスクのために拡張し、分布類似度によって相互情報量を重み付けする。

分布類似度による言い換え獲得では、同義語と反義語や類義語の判別が難しい [7] という問題があるが、Bilingual Pivoting のような単語アライメント誤りに起因する関連のない語句によるノイズは起こりにくい。また、Bilingual Pivoting には分布類似度のような反義語や類義語によるノ

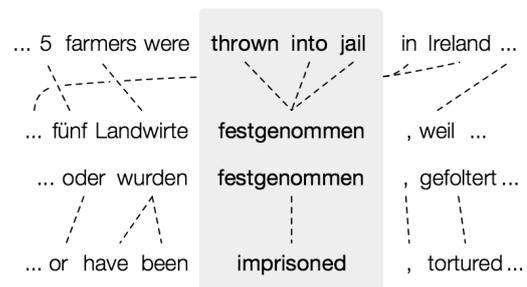


図 1 Bilingual Pivoting による言い換え獲得 [1]

イズは起こりにくい。そのため、Bilingual Pivoting と分布類似度を組み合わせる提案手法には、従来の言い換え獲得手法の欠点を補い合う高い頑健性が期待できる。

実験の結果、我々が提案する分布類似度を用いて重み付けした相互情報量に基づく言い換え獲得手法は、MRR と MAP の両方で Bilingual Pivoting および分布類似度を改善できた。本研究の貢献は以下の 4 点である。

- 単語間の言い換え確率を相互情報量を用いて一般化した。
- 対訳コーパス (Bilingual Pivoting) と単言語コーパス (単語出現確率および分布類似度) の両方を用いて高精度に語彙的言い換えを獲得した。
- Kneser-Ney Smoothing を用いて Bilingual Pivoting の過推定を抑制した。
- 獲得した英語と日本語の言い換え対を公開^{*1}した。

¹ 首都大学東京 システムデザイン研究科

² 統計数理研究所

a) kajiwara-tomoyuki@ed.tmu.ac.jp

b) komachi@tmu.ac.jp

c) daichi@ism.ac.jp

^{*1} <https://github.com/tmu-nlp/pmi-ppdb>

2. Bilingual Pivoting

Bilingual Pivoting [5] では、対訳コーパス中で英語のフレーズ e_1 と外国語のフレーズ f との間に単語アライメントがあり、外国語のフレーズ f と英語のフレーズ e_2 の間にも単語アライメントがあるとき、 e_1 から e_2 への条件付き言い換え確率 $p(e_2|e_1)$ を、単語アライメント確率 $p(f|e_1)$ および $p(e_2|f)$ を用いて次のように仮定する。

$$\begin{aligned} p(e_2|e_1) &= \sum_f p(e_2|f, e_1)p(f|e_1) \\ &\approx \sum_f p(e_2|f)p(f|e_1) \end{aligned} \quad (1)$$

PPDB [1] では、式 (1) の Bilingual Pivoting を用いて、対称な言い換えスコア $s_{bp}(e_1, e_2)$ を次のように定義する。ただし、本研究では $\lambda_1 = \lambda_2 = -1^*$ とする。

$$\begin{aligned} s_{bp}(e_1, e_2) &= -\lambda_1 \log p(e_2|e_1) - \lambda_2 \log p(e_1|e_2) \\ &= \log p(e_2|e_1) + \log p(e_1|e_2) \end{aligned} \quad (2)$$

これらの Bilingual Pivoting による言い換え獲得は、大規模な一方で、単語アライメント誤りに起因するノイズも多い。そこで本研究では、Kneser-Ney Smoothing [8] を用いた過推定のスムージング (3 節) および大規模な単言語コーパスから得られる単語出現頻度を用いたラランキンク (4 節) によって、高精度な言い換え獲得を行う。

3. Bilingual Pivoting における過推定の抑制

言い換え確率の推定に利用する対訳コーパスのデータスパースネス問題のために、低頻度の単語対に対して言い換え確率を過推定してしまうことがある。Bilingual Pivoting では、単語間の言い換え確率を条件付き確率 $p(e_2|e_1)$ で仮定するため、条件付き確率のスムージング手法を適用することによってこの過推定の問題を抑制することができる。

階層ベイズモデルでは、ディリクレ分布 α_y を仮定して条件付き確率 $p(y|x)$ を最尤推定値 $\hat{p}_{y|x}$ を用いて次のように表現する。ただし、 $n(x)$ は単語 x の出現頻度である。

$$\begin{aligned} p(y|x) &= \frac{n(y|x) + \alpha_y}{\sum_y (n(y|x) + \alpha_y)} \\ &= \frac{n(y|x)}{n(x) + \sum_y \alpha_y} \quad \because \alpha_y \ll 1 \\ &= \frac{n(x)}{n(x) + \sum_y \alpha_y} \cdot \frac{n(y|x)}{n(x)} \\ &= \frac{n(x)}{n(x) + \sum_y \alpha_y} \cdot \hat{p}_{y|x} \end{aligned} \quad (3)$$

この $\sum_y \alpha_y$ は無視できない値なので、特に頻度 $n(x)$ が小さいとき、式 (3) は最尤推定 $\hat{p}_{y|x}$ が確率を過剰に大きく推定してしまうことを意味している。

*2 公開されている PPDB*3 では $\lambda_1 = \lambda_2 = 1$

*3 <http://www.cis.upenn.edu/~ccb/ppdb/>

本研究では、N-gram 言語モデルにおける条件付き確率のスムージング手法としてよく知られている Kneser-Ney Smoothing [8] を用いて、Bilingual Pivoting における言い換え確率の過推定を次のように抑制する。

$$\begin{aligned} p_{kn}(e_2|e_1) &= \frac{n(e_2|e_1) - \delta}{n(e_1)} + \gamma(e_1)p_{kn}(e_2) \\ \delta &= \frac{N_1}{N_1 + 2N_2} \\ \gamma(e_1) &= \frac{\delta}{n(e_1)} N(e_1) \\ p_{kn}(e_2) &= \frac{N(e_2)}{\sum_i N(e_i)} \end{aligned} \quad (4)$$

ここで、 N_n は n 回だけ出現する単語対の種類数、 $N(e_1)$ は単語 e_1 の言い換え候補の種類数をそれぞれ表す。

4. 相互情報量に基づく言い換え獲得手法

4.1 相互情報量に基づく Bilingual Pivoting の一般化

本研究では、対訳コーパスよりも十分に大きい単言語コーパスから得られる単語出現確率 $p(e_1)$ および $p(e_2)$ を用いて、式 (2) の Bilingual Pivoting による言い換えらしさを次のようにスムージングする。

$$\begin{aligned} s_{pmi}(e_1, e_2) &= \log p(e_2|e_1) + \log p(e_1|e_2) \\ &\quad - \log p(e_1) - \log p(e_2) \end{aligned} \quad (5)$$

対数の差は商の対数に等しいので、式 (5) は次のように変形でき、言い換えらしさを自己相互情報量で説明できる。

$$\begin{aligned} s_{pmi}(e_1, e_2) &= \log \frac{p(e_2|e_1)}{p(e_2)} + \log \frac{p(e_1|e_2)}{p(e_1)} \\ &= 2\text{PMI}(e_1, e_2) \end{aligned} \quad (6)$$

なぜならば、自己相互情報量 $\text{PMI}(x, y)$ はベイズの定理によって条件付き確率を用いる次の形に変形できる。

$$\begin{aligned} \text{PMI}(x, y) &= \log \frac{p(x, y)}{p(x)p(y)} \\ &= \log \frac{p(y|x)p(x)}{p(x)p(y)} = \log \frac{p(y|x)}{p(y)} \\ &= \log \frac{p(x|y)p(y)}{p(x)p(y)} = \log \frac{p(x|y)}{p(x)} \end{aligned} \quad (7)$$

そのため、 $p(x, y) = p(y|x)p(x)$ と $p(x, y) = p(x|y)p(y)$ の両方を用いた幾何平均によって、式 (6) が導出できる。

$$\begin{aligned} \text{PMI}(x, y) &= \frac{1}{2}\text{PMI}(x, y) + \frac{1}{2}\text{PMI}(x, y) \\ &= \frac{1}{2} \log \frac{p(y|x)}{p(y)} + \frac{1}{2} \log \frac{p(x|y)}{p(x)} \\ &= \log \left[\left\{ \frac{p(y|x)}{p(y)} \right\}^{\frac{1}{2}} \cdot \left\{ \frac{p(x|y)}{p(x)} \right\}^{\frac{1}{2}} \right] \end{aligned} \quad (8)$$

通常、自然言語処理における自己相互情報量は、単語の共起確率を用いて単語間の関連の強さを表す。しかし、本

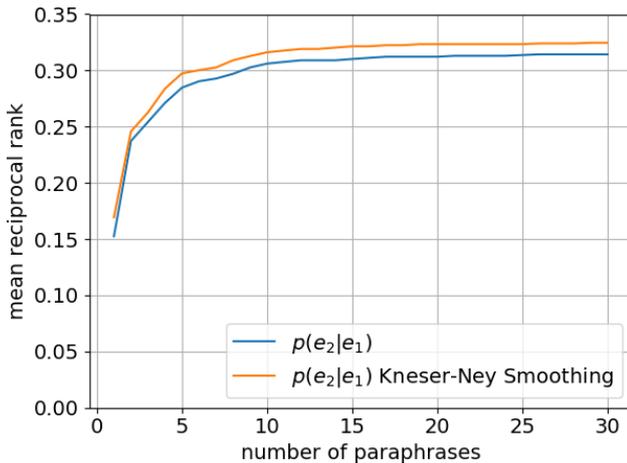


図 2 MRR: Kneser-Ney Smoothing の有効性

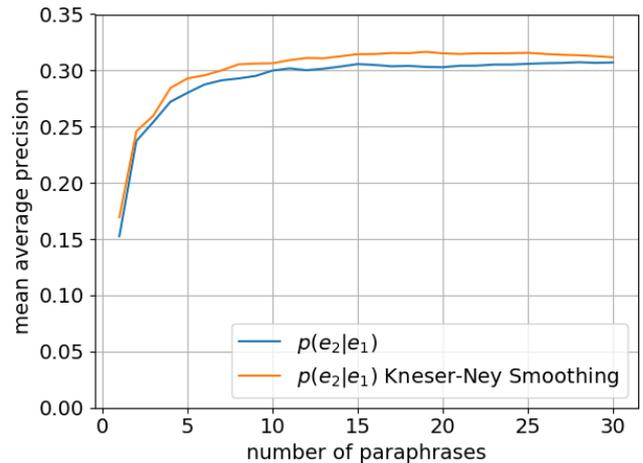


図 3 MAP: Kneser-Ney Smoothing の有効性

研究では自己相互情報量によって Bilingual Pivoting の言い換え確率を一般化するので、式 (6) における $p(e_2|e_1)$ はある文に e_1 が出現した場合に同じ文に e_2 が出現する条件付き共起確率ではなく、式 (1) の条件付き言い換え確率 (対訳コーパス上で周辺化された単語アライメント確率) であることに注意されたい。

式 (1) の Bilingual Pivoting は $e_1 \rightarrow e_2$ 方向のみを考慮する Mixture Model と見なすことができるが、式 (8) に示したように我々の提案手法は両方向を考慮する Product Model [9] と見なすことができる。PPDB [1, 2] も両方向の言い換え確率を考慮しているが、彼らはこれを Product Model とは考えておらず、各方向の言い換え確率を教師あり学習の素性のひとつとして扱っている。我々の提案手法では相互に置換可能な単語対のみが高いスコアを持つことになるが、これは言い換えの双方向性を踏まえると妥当なモデル化であると考えられる。

4.2 相互情報量における低頻度バイアス問題への対処

低頻度な単語対では、偶然の共起により相互情報量が不当に大きくなる低頻度バイアスの問題 [10] がよく知られている。この問題に対処するため、単語対の共起頻度によって重み付けを行う LocalPMI [6] が提案されている。

$$\text{LocalPMI}(x, y) = n(x, y) \cdot \text{PMI}(x, y) \quad (9)$$

本研究では、式 (9) の $n(x, y)$ に当たる重みを対訳コーパス上で直接求めることは難しい。また、本研究で求めたいものは単語間の共起 (関連) の強さではなく、単語間の言い換えらしさであるため、一般的な LocalPMI のように単言語コーパス上で窓枠を設定して共起頻度を数えることも適切ではない。

そこで我々は、単言語コーパスを用いた言い換え獲得としてよく利用される分布類似度 [11, 12] を重みとして、LocalPMI を言い換え獲得のために次のように拡張する。

$$\begin{aligned} s_{lpmi}(e_1, e_2) &= \cos(e_1, e_2) \cdot s_{pmi}(e_1, e_2) \\ &= \cos(e_1, e_2) \cdot 2\text{PMI}(e_1, e_2) \end{aligned} \quad (10)$$

ここで、 $\cos(e_1, e_2)$ は、単語 e_1 および単語 e_2 のベクトル表現を用いた余弦類似度である。この式 (10) は、単言語コーパスに基づく言い換えらしさ (分布類似度) と式 (6) の対訳コーパスに基づく言い換えらしさ (Bilingual Pivoting) を同時に考慮することを意味する。分布類似度では Bilingual Pivoting のような関連のない語句によるノイズは起こりにくい。また、Bilingual Pivoting では分布類似度のような反義語や類義語によるノイズは起こりにくい。そのため、本手法では従来手法における欠点を互いに補う頑健な言い換え獲得が期待できる。

5. 実験

5.1 相互情報量に基づく言い換え獲得

対訳コーパスには Europarl-v7 [13] の英仏データ*4 を利用し、GIZA++ [14] (IBM model 4) によって単語アライメント確率 $p(e_2|e_1)$ および $p(e_1|e_2)$ を計算した。単言語コーパスには English Gigaword 5th Edition*5 を利用し、KenLM [15] によって $p(e_1)$ および $p(e_2)$ を計算した。 $\cos(e_1, e_2)$ には、word2vec [16] の cbow モデル*6 を用いた。

まず、英仏対訳コーパスを用いて、英→仏および仏→英の両方向で単語アライメント確率を計算した。そして、Kneser-Ney Smoothing を用いて過推定を抑制した言い換え確率を求めた。この言い換え確率と、単言語コーパスから得られる単語出現確率を用いて相互情報量に基づく言い換えスコアを計算し、低頻度バイアスを避けるために余弦類似度による重み付けを行った。

最終的に、自分自身への言い換え ($e_1 = e_2$) を除き、170,682,871 単語対の言い換え候補を得た。

*4 <http://www.statmt.org/europarl/>

*5 <https://catalog.ldc.upenn.edu/LDC2011T07>

*6 <https://code.google.com/archive/p/word2vec/>

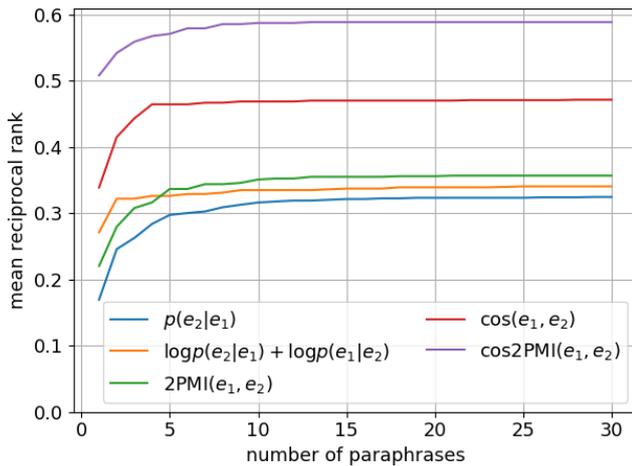


図 4 MRR: 相互情報量に基づく言い換え獲得

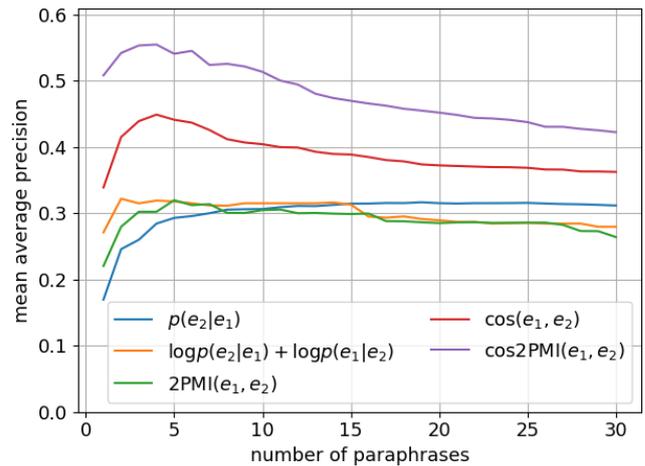


図 5 MAP: 相互情報量に基づく言い換え獲得

5.2 言い換え対の評価: MRR および MAP

Pavlick et al. [2] と同じく、Mean Reciprocal Rank (MRR) および Mean Average Precision (MAP) によって言い換え対を評価する。Pavlick et al. [2] によって公開されている Human Paraphrase Judgments^{*7}には、Wikipedia から無作為抽出された 100 単語についての言い換えリストおよび各言い換え対に対する 5 段階の人手評価が含まれている。本節では、このデータを用いて、獲得した言い換え対の評価を行う。ただし、正解の言い換えは人手評価において 5 段階評価の 3 以上の評価を得た単語のみとする。

図 2 および図 3 に、Bilingual Pivoting に Kneser-Ney Smoothing を加えたときの MRR および MAP の変化を示す。各グラフの横軸は、言い換えスコアの上位 k 番目までの言い換えを評価することを表す。図 2 および図 3 より、Kneser-Ney Smoothing によって Bilingual Pivoting で獲得した言い換え候補のランキングが改善されることがわかる。以下、本稿では Bilingual Pivoting には常に Kneser-Ney Smoothing を適用して実験する。

続いて図 4 および図 5 に、提案手法である相互情報量に基づく言い換え獲得の MRR および MAP による評価を示す。グラフはそれぞれ、青線が Bilingual Pivoting (ベースライン)、黄線が PPDB における対称な言い換えスコア (ベースライン)、緑線が相互情報量を用いる言い換えスコア (提案手法)、赤線が分布類似度 (ベースライン)、紫線が分布類似度によって重み付けした相互情報量を用いる言い換えスコア (提案手法) である。赤線の分布類似度とは、全語彙の中から余弦類似度の上位 k 単語を利用したのではなく、Bilingual Pivoting によって獲得された言い換え候補を余弦類似度によってリランキングしたということに注意されたい。図 4 の MRR による評価に着目すると、まず Bilingual Pivoting は対称化することによって言い換

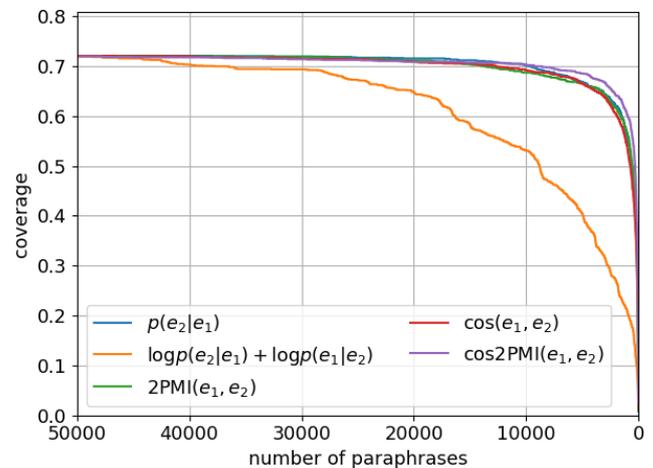


図 6 上位 k 番目までの言い換への網羅性

え候補のランキング性能が改善される。提案手法である相互情報量を用いる言い換えスコアは、上位 5 番目以降では Bilingual Pivoting のベースラインをわずかに上回る。そして、分布類似度による重み付けにより、相互情報量を用いる言い換えスコアは最も高い性能を示した。なお、分布類似度によるリランキングは、単体でも Bilingual Pivoting を大きく改善しており、対訳コーパスから得られる情報に加えて単言語コーパスから得られる情報を利用することの重要性がわかる。図 5 の MAP による評価も同様で、分布類似度によるリランキングが Bilingual Pivoting を改善し、提案手法である分布類似度により重み付けした相互情報量が最も高い性能を示した。

5.3 言い換え対の評価: Coverage

5.2 節と同じデータを用いて、上位 k 番目までの言い換え対における網羅性を評価する。as などの機能語は、Bilingual Pivoting における単語アライメント誤りの影響を受けやすく、50,000 種類を超える非常に多くの言い換え

*7 <http://www.seas.upenn.edu/~epavlick/data.html>

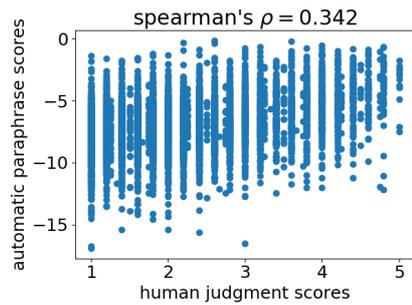


図 7 $\rho : p(e_2|e_1)$

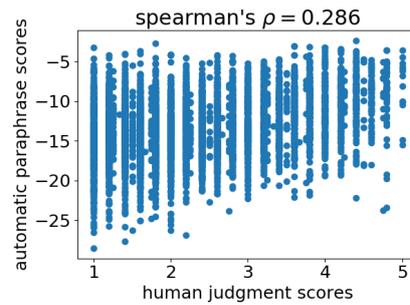


図 8 $\rho : \log p(e_2|e_1) + \log p(e_1|e_2)$

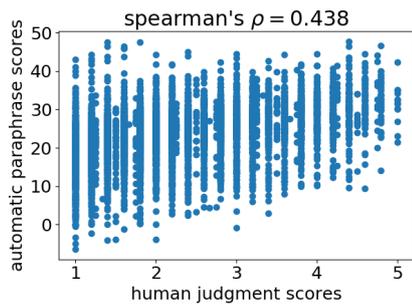


図 9 $\rho : 2PMI(e_1, e_2)$

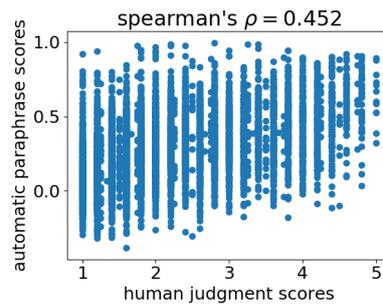


図 10 $\rho : \cos(e_1, e_2)$

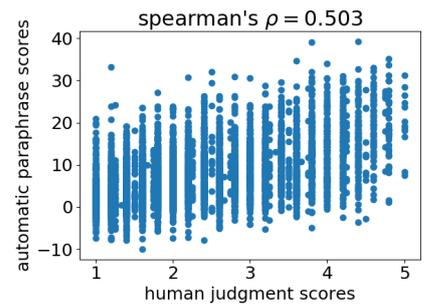


図 11 $\rho : \cos(e_1, e_2)2PMI(e_1, e_2)$

候補を持つ。しかし、これらの言い換え候補の多くは実際には言い換えの関係ではない単語対であるため、正解の言い換えを保持しつつ不要な候補をいかに削減できるかという網羅性の評価を行う。

図 6 に網羅性の評価を示す。黄線の対称化された Bilingual Pivoting は、5.2 節の実験結果では上位のランキング性能において $p(e_2|e_1)$ や重み付けをしない相互情報量よりも優れていたにも関わらず、正解の言い換えを下位にも含んでしまっている。対称化された Bilingual Pivoting 以外は、いずれの言い換えランキング手法を用いても網羅性については大差がないが、紫線の分布類似度による重み付けを行った相互情報量は、他の手法よりもわずかに良い。

5.4 言い換え対の評価：相関係数

MRR および MAP による上位の言い換え対の評価、Coverage による下位のランキングの評価に加えて、相関係数によって全体的な言い換えランキングの妥当性を評価する。我々は Pavlick et al. [2] と同じく、スピアマンの順位相関係数によって言い換え対を評価する。Pavlick et al. [2] によって公開されている Human Paraphrase Judgments^{*4}には、前節で利用した言い換えリストの他に、PPDB [1] からサンプリングされた 26,456 単語対に対する 5 段階の人手評価も含まれている。本節では、この後者のデータを用いて、獲得した言い換え対を評価する。

図 7 から図 11 に、各言い換えランキングスコアと人手評価（5 人の評価の平均値）の散布図およびスピアマンの

順位相関係数を示す。5.3 節で示した通り、PPDB における対称な Bilingual Pivoting は上位のスコア以外は信頼性に欠けるため、相関係数も低い。提案手法である分布類似度により重み付けした相互情報量は、相関係数においても最も高い性能を示した。特に散布図の左上のノイズ (false positive) が削減できていることがわかる。

6. 考察

6.1 言い換え対の定性評価

表 1 に、言い換えランキングの上位の例を示す。cultural の言い換えの例では、Bilingual Pivoting は、正しい言い換えが上位に出現せず、上位の単語は cultural とは関連が弱い。PPDB における対称な言い換えスコアは、トップに正しい言い換えを獲得できているものの、これまでも示したようにトップ以外の単語は信頼できない。相互情報量は、低頻度語の影響を強く受けており、上位の単語の多くが対訳コーパス中での出現頻度 1 の単語である。提案手法である分布類似度により重み付けした相互情報量は、相互情報量における低頻度バイアスの問題が解決され、上位の単語の多くが正しい言い換えである。分布類似度を利用する 2 手法は、比較的多くの正しい言い換えを上位に持ち、上位の他の単語も cultural との関連が強い。言い換えという観点では提案手法の上位 10 単語のうち 3 単語は不正解であるが、意味的文間類似度 [3] や単語分散表現の学習 [4] など、これまでに PPDB が利用されている応用タスクの一部ではこれらも有用である可能性がある。

表 1 *cultural* の言い換えランキングの上位 (斜体が正しい言い換え)

$p(e_2 e_1)$	$\log p(e_2 e_1) + \log p(e_1 e_2)$	2PMI(e_1, e_2)	$\cos(e_1, e_2)$	$\cos(e_1, e_2)2PMI(e_1, e_2)$
1. diverse	<i>culturally</i>	culturally-based	historical	<i>socio-cultural</i>
2. harvests	<i>culture</i>	culturaldevelopment	<i>culture</i>	<i>culture</i>
3. firstly	151	cultural-social	educational	<i>multicultural</i>
4. understand	charter	economic-cultural	linguistic	<i>intercultural</i>
5. flowering	monuments	culture-	<i>multicultural</i>	educational
6. trying	art	cultural-educational	<i>cross-cultural</i>	intellectual
7. structure	casal	kulturkampf	diversity	<i>culturally</i>
8. january	kahn	cultural-political	technological	<i>sociocultural</i>
9. <i>culture</i>	13	multiculture	intellectual	<i>heritage</i>
10. <i>culturally</i>	caning	<i>culturally</i>	preservation	architectural

表 2 *labourers* の言い換えランキング (正しい言い換えのみ)

$p(e_2 e_1)$	$\log p(e_2 e_1) + \log p(e_1 e_2)$	2PMI(e_1, e_2)	$\cos(e_1, e_2)$	$\cos(e_1, e_2)2PMI(e_1, e_2)$
1. workers	9. gardeners	10. workmen	2. workers	2. workers
2. employees	42. harvesters	11. wage-earners	8. people	4. workmen
9. farmers	62. workers	16. earners	10. persons	5. craftsmen
13. labour	71. seafarers	19. workers	11. farmers	6. wage-earners
16. gardeners	73. unions	21. craftsmen	15. craftsmen	9. persons
17. people	99. homeworkers	22. workforces	26. wage-earners	12. employees
28. workmen	283. works	26. employed	27. workmen	13. earners
30. employed	394. workmen	27. employees	29. harvesters	15. farmers
33. craftsmen	395. employees	50. labour	31. seafarers	18. people
59. harvesters	412. wage-earners	55. persons	32. employees	19. workforces
80. work	415. craftsmen	75. farmers	42. gardeners	37. harvesters
88. earners	417. earners	103. homeworkers	47. earners	42. individuals
90. wage-earners	419. labour	105. individuals	55. workforces	53. labour
106. persons	420. employed	112. work	57. individuals	55. seafarers
109. individuals	431. people	135. people	79. unions	65. gardeners
114. seafarers	433. farmers	187. harvesters	103. labour	88. employed
115. unions	446. workforces	273. gardeners	140. homeworkers	100. homeworkers
131. workforces	451. work	317. seafarers	144. work	105. work
166. homeworkers	453. persons	456. unions	170. employed	149. unions
401. works	474. individuals	469. works	222. works	254. works

表 2 に、言い換えランキングのうちの正しい言い換えとそのランキングを示す。*labourers* の言い換える例では、人手評価において 5 段階評価の 3 以上の評価を得た正しい言い換えが 20 語存在した。Bilingual Pivoting や相互情報量では、全ての正しい言い換えを網羅するためには上位 400 位まで考慮する必要がある。一方で、分布類似度を利用する手法では、正しい言い換えを比較的上位に持つことができている。特に、提案手法である分布類似度により重み付けした相互情報量では、上位 20 単語までに 10 単語の正しい言い換えを含めることができた。すなわち、提案手法では上位の言い換えを使用するだけでカバレッジの高い言い換えを得ることができる。

6.2 言い換え対の外的評価

意味的文間類似度タスクで言い換え対の外的評価を行う。意味的文間類似度タスクは 2 文間の意味的な類似度を求めるタスクであり、本研究では SemEval [17-21] で

使用された 5 つのデータセットを用いて 5 段階の人手評価との相関 (ピアソンの積率相関) を評価する。我々は、SemEval-2015 の意味的文間類似度タスク [20] で優秀な成績を収めた DLS@CU [3] の教師なし手法に対して、本研究で獲得した言い換え対を適用する。DLS@CU では PPDB を用いた単語アライメント [22] を行い、アラインされた単語の割合に応じて式 (11) のように文間類似度を決定する。

$$sts(s_1, s_2) = \frac{n_a(s_1) + n_a(s_2)}{n(s_1) + n(s_2)} \quad (11)$$

ここで、 $n(s)$ は文 s の単語数、 $n_a(s)$ はアラインされた単語数を意味する。DLS@CU では PPDB の全ての言い換えを対象にしているが、本研究では各単語に対して言い換えスコアの上位 10 単語のみを使用し、言い換えスコアの性能を比較する。

表 3 に、実験結果を示す。ALL とは、5 つのデータセットにおけるピアソン相関係数の重み付き平均値である。提案手法が 5 つのデータセットのうち 3 つで最高性能を達成

表 3 意味的文間類似度タスクへの応用 (ピアソンの積率相関による評価)

	$p(e_2 e_1)$	$\log p(e_2 e_1) + \log p(e_1 e_2)$	$2\text{PMI}(e_1, e_2)$	$\cos(e_1, e_2)$	$\cos(e_1, e_2)2\text{PMI}(e_1, e_2)$
STS-2012	0.539	0.466	0.383	0.363	0.442
STS-2013	0.489	0.469	0.463	0.483	0.499
STS-2014	0.464	0.460	0.471	0.453	0.475
STS-2015	0.611	0.655	0.660	0.642	0.671
STS-2016	0.444	0.518	0.550	0.518	0.542
ALL	0.536	0.543	0.534	0.523	0.555

しており、文間類似度計算のために有用な言い換え対を上位にランキングできていることがわかる。

7. 関連研究

Levy and Goldberg [23] は、単語分散表現の学習手法としてよく知られている Mikolov et al. [16] の skip-gram with negative-sampling (SGNS) を Shifted Positive PMI の行列分解として説明している。本研究は、言い換え獲得手法としてよく知られている Bannard and Callison-Burch [5] の Bilingual Pivoting を PMI によって説明するものである。

Bhagat and Ravichandran [24] は、言い換え獲得のために PMI を利用している。しかし、彼らは窓幅 1 の共起語 c を PMI で重み付けした単語ベクトル V 間の余弦類似度によって、単語間の意味的な類似度を計算している。

$$\begin{aligned} \text{PMI}(e_i, c) &= \log \frac{p(e_i, c)}{p(e_i)p(c)} \\ \cos(e_i, e_j) &= \frac{V_i \cdot V_j}{|V_i||V_j|} \end{aligned} \quad (12)$$

彼らは単言語コーパス上で言い換え候補と共起語との共起頻度を用いて PMI を計算しているのに対して、我々是对訳コーパス上で言い換え候補同士の単語アライメント確率を用いて PMI を計算していることが異なる。

Chan et al. [11] は、Bilingual Pivoting で獲得された言い換え対を分布類似度によってリランキングしている。単言語コーパスから得られる情報を用いて言い換えをリランキングするという考え方は我々と似ているが、彼らの手法では意味的に近い言い換えを集めることができていない。本研究では LocalPMI による定式化によって、対訳コーパスから得られる情報と単言語コーパスから得られる情報を効果的に組み合わせ、良い言い換えを集めることができた。

Bilingual Pivoting [5] を用いて、多くの言語で言い換えデータベースが構築されている。Ganitkevitch and Callison-Burch [25] は欧州言語や中国語など 23 言語の言い換えデータベース^{*8}を構築し、Mizukami et al. [26] は日本語の言い換えデータベース^{*9}を構築した。本研究では、単言語コーパスを用いて Bilingual Pivoting を改善した。大規模な単言語コーパスは多くの言語で容易に利用できるため、これらの各言語で言い換え知識を改善できるだろう。

^{*8} <http://paraphrase.org/>

^{*9} <http://ahclab.naist.jp/resource/jppdb/>

Bilingual Pivoting によって構築された PPDB [1] は、意味的文間類似度 [3]、単語分散表現の学習 [4]、機械翻訳 [27]、文圧縮 [28]、質問応答 [29]、テキスト平易化 [30] など、多くの自然言語処理応用タスクで利用されている。本研究で提案した Bilingual Pivoting の改善手法は、PPDB を利用するこれらの多くのタスクの性能を改善する可能性がある。

8. おわりに

本研究では、単語間の言い換えらしさを重み付き相互情報量によって定義し、対訳コーパスから得られる情報と単言語コーパスから得られる情報の両方を利用して言い換え対を獲得した。分布類似度で重み付けした Bilingual Pivoting は、それぞれの言い換え獲得手法が相補的にはたらくことにより、頑健に言い換え獲得ができた。人手でアノテーションされた言い換え対の評価用データセットを用いた実験の結果、MRR や MAP において従来の Bilingual Pivoting や分布類似度に対する性能の改善が確認できた。また、意味的文間類似度タスクにおける外的評価からも、提案手法の有効性が確認できた。本研究で扱った意味的文間類似度以外にも、機械翻訳や文圧縮、質問応答やテキスト平易化など、語彙的言い換え知識は多くのタスクで利用されているため、本研究の成果をもとに多くの自然言語処理応用タスクの性能を改善できると期待する。

参考文献

- [1] Ganitkevitch, J., Van Durme, B. and Callison-Burch, C.: PPDB: The Paraphrase Database, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 758–764 (2013).
- [2] Pavlick, E., Rastogi, P., Ganitkevitch, J., Van Durme, B. and Callison-Burch, C.: PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pp. 425–430 (2015).
- [3] Sultan, M. A., Bethard, S. and Sumner, T.: DLS@CU: Sentence Similarity from Word Alignment and Semantic Vector Composition, *Proceedings of the 9th International Workshop on Semantic Evaluation*, pp. 148–153 (2015).
- [4] Yu, M. and Dredze, M.: Improving Lexical Embeddings with Semantic Knowledge, *Proceedings of the 52nd Annual Meeting of the Association for Computational Lin-*

- guistics*, pp. 545–550 (2014).
- [5] Bannard, C. and Callison-Burch, C.: Paraphrasing with Bilingual Parallel Corpora, *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pp. 597–604 (2005).
- [6] Evert, S.: The Statistics of Word Cooccurrences: Word Pairs and Collocations, PhD Thesis, University of Stuttgart (2005).
- [7] Mohammad, S. M., Dorr, B. J., Hirst, G. and Turney, P. D.: Computing Lexical Contrast, *Computational Linguistics*, Vol. 39, No. 3, pp. 555–590 (2013).
- [8] Kneser, R. and Ney, H.: Improved Backing-off for M-gram Language Modeling, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. 1, pp. 181–184 (1995).
- [9] Hinton, G. E.: Training Products of Experts by Minimizing Contrastive Divergence, *Neural Computation*, Vol. 14, No. 8, pp. 1771–1800 (2002).
- [10] Levy, O., Goldberg, Y. and Dagan, I.: Improving Distributional Similarity with Lessons Learned from Word Embeddings, *Transactions of the Association for Computational Linguistics*, Vol. 3, pp. 211–225 (2015).
- [11] Chan, T. P., Callison-Burch, C. and Van Durme, B.: Reranking Bilingually Extracted Paraphrases Using Monolingual Distributional Similarity, *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pp. 33–42 (2011).
- [12] Glavaš, G. and Štajner, S.: Simplifying Lexical Simplification: Do We Need Simplified Corpora?, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pp. 63–68 (2015).
- [13] Koehn, P.: Europarl: A Parallel Corpus for Statistical Machine Translation, *Proceedings of the Machine Translation Summit*, pp. 79–86 (2005).
- [14] Och, F. J. and Ney, H.: A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, Vol. 29, No. 1, pp. 19–51 (2003).
- [15] Heafield, K.: KenLM: Faster and Smaller Language Model Queries, *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pp. 187–197 (2011).
- [16] Mikolov, T., Chen, K., Corrado, G. S. and Dean, J.: Efficient Estimation of Word Representations in Vector Space, *Proceedings of Workshop at the International Conference on Learning Representations*, pp. 1–12 (2013).
- [17] Agirre, E., Cer, D., Diab, M. and Gonzalez-Agirre, A.: SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity, **SEM 2012: The First Joint Conference on Lexical and Computational Semantics*, pp. 385–393 (2012).
- [18] Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A. and Guo, W.: *SEM 2013 shared task: Semantic Textual Similarity, *Second Joint Conference on Lexical and Computational Semantics*, pp. 32–43 (2013).
- [19] Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Mihalcea, R., Rigau, G. and Wiebe, J.: SemEval-2014 Task 10: Multilingual Semantic Textual Similarity, *Proceedings of the 8th International Workshop on Semantic Evaluation*, pp. 81–91 (2014).
- [20] Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Lopez-Gazpio, I., Maritxalar, M., Mihalcea, R., Rigau, G., Uria, L. and Wiebe, J.: SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability, *Proceedings of the 9th International Workshop on Semantic Evaluation*, pp. 252–263 (2015).
- [21] Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R., Rigau, G. and Wiebe, J.: SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation, *Proceedings of the 10th International Workshop on Semantic Evaluation*, pp. 497–511 (2016).
- [22] Sultan, M. A., Bethard, S. and Sumner, T.: Back to Basics for Monolingual Alignment: Exploiting Word Similarity and Contextual Evidence, *Transactions of the Association for Computational Linguistics*, Vol. 2, pp. 219–230 (2014).
- [23] Levy, O. and Goldberg, Y.: Neural Word Embedding as Implicit Matrix Factorization, *Advances in Neural Information Processing Systems*, pp. 2177–2185 (2014).
- [24] Bhagat, R. and Ravichandran, D.: Large Scale Acquisition of Paraphrases for Learning Surface Patterns, *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 674–682 (2008).
- [25] Ganitkevitch, J. and Callison-Burch, C.: The Multilingual Paraphrase Database, *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pp. 4276–4283 (2014).
- [26] Mizukami, M., Neubig, G., Sakti, S., Toda, T. and Nakamura, S.: Building a Free, General-Domain Paraphrase Database for Japanese, *Proceedings of the 17th Oriental COCOSDA Conference*, pp. 129 – 133 (2014).
- [27] Mehdizadeh Seraj, R., Siabani, M. and Sarkar, A.: Improving Statistical Machine Translation with a Multilingual Paraphrase Database, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1379–1390 (2015).
- [28] Napoles, C., Callison-Burch, C. and Post, M.: Sentential Paraphrasing as Black-Box Machine Translation, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 62–66 (2016).
- [29] Sultan, M. A., Castelli, V. and Florian, R.: A Joint Model for Answer Sentence Ranking and Answer Extraction, *Transactions of the Association for Computational Linguistics*, Vol. 4, pp. 113–125 (2016).
- [30] Xu, W., Napoles, C., Pavlick, E., Chen, Q. and Callison-Burch, C.: Optimizing Statistical Machine Translation for Text Simplification, *Transactions of the Association for Computational Linguistics*, Vol. 4, pp. 401–415 (2016).